データベースからデータへ

関野　樹

昨秋、日文研において、ＩＩＩＦ[註一]に関するワークショップが開催された。ＩＩＩＦはInternational Image Interoperability Framework の略で、画像データを相互運用するための国際的な規格である。この規格に則って公開された画像データは、どこから公開されたかに関わらず、同じ手順で利用することができる。利用者は、自身が使い慣れた閲覧用ソフトウェアを使って、さまざまな機関のデータベースの画像を閲覧することができ、データベースごとに閲覧操作を覚える必要はない。さらに、画像の注目する部分を切り抜いて表示したり、それらを相互に比較したりすることも可能である。国立国会図書館をはじめとする多くの機関でＩＩＩＦによる画像データの公開が進められており、関連学会でも、ＩＩＩＦの仕組みを応用したデータの利活用に関する提案が数多くなされている。今回のワークショップにも、近隣の大学附属図書館の担当者など、センター外からの参加があり、関心の高さが窺える。

ＩＩＩＦによる画像データの公開では、データを提供する側と利用する側の立場や役割が従来とは異なってくる。これまでであれば、利用者はまずＷｅｂブラウザでデータベースにアクセスし、キーワードなどを使って必要なデータを検索する。その上で、得られたデータをＷｅｂブラウザ上で閲覧するという手順でデータを利用してきた。検索、閲覧とも、データを提供する側が用意した機能を利用しており、提供者の意図した方法、言うなれば、見せたい形で

データを検索・閲覧してもらうことができる。

一方、ＩＩＩＦで公開された画像データは、閲覧方法も含め、データの利活用は利用者に委ねられる。個々のデータはデータベースという枠を離れ、それぞれ独立した研究資源として扱われる。このため、データベースは、データに辿り着くための入口の一つであり、また、そこで提供されている閲覧機能も、データ利活用の一例という位置付けとなる。ＩＩＩＦは、付随するデータ（メタデータ）やアクセス手段を標準化することにより、個別の画像データを独立した研究資源として扱えるようにするための規格と捉えることもできる。

データそのものを直接利用しようとする動きは、画像データだけでなく、そのほかのデータにも拡がりを見せている。その一例として、ここ数年で普及してきたリンクト・データ（Linked Data）がある。これは、Ｗｅｂページのリンクのように、データを相互に連携した形で公開するための仕組みである（「データのＷｅｂ」とも言われる）。提供元やデータベースの違いを超えて連携した多種多様なデータは、データの検索や解析に用いられるほか、さまざまな事象の識別や典拠を示すことなどにも利用される。たとえば、日付に関するリンクト・データでは[註二]、日本の明治以前の日付も含めて、各国のさまざまな暦の日付とリンクした形でデータが提供されている。この日付に関するリンクト・データを仲立ちにすることで、暦の違いを超えて、さまざまなデータを時間という点で相互に関連付けて利用することが可能になる。リンクト・データによるデータの公開は、研究機関だけでなく、行政機関でも試みが進んでおり、今後の拡がりが期待されている。

これらのデータそのものを利活用しようとする動きの背景として、データの利用方法が多様化したことが挙げられる。画像などのデータを画面上で閲覧するだけではなく、独自の解析や

比較を行ったり、データを使った新たなソフトウェアやサービスを開発したりといったことが行われるようになっている。ＡＩ（人工知能）への応用は、その典型例である。さまざまなタイプのＡＩがあるが、最近注目を集めているのが深層学習（ディープ・ラーニング）を含む、機械学習の手法を応用したものである。これは、目的とする処理の判断材料となるデータをコンピュータに学習させ、その学習結果に基づいて実際の処理を行う手法である。たとえば、ＡＩを使ってくずし字を自動認識させる取り組みが進められている。[註三] これを実現するためには、さまざまな史資料の画像から文字の見本を集める必要があり、これらの見本を使って学習した結果に基づいて、自動認識の処理が実現する。ここで使われている史資料画像の多くは、内容を読んでもらうことを想定して公開されたものであろうが、良い意味で「想定外」のデータ利活用が実現したことになる。

利用者によるデータの自由な利活用は、一定のリスクを伴うことも事実である。たとえば、データ提供者の本意ではない使われ方をするのではといったことは、誰もが危惧するところである。また、データの利用状況が把握しにくくなることも、研究評価の面で気になってくる。その反面、上述のように、自由な利用が提供者の予期しない新たな発見や研究ツールを生み出す可能性も秘めている。こうした成果は、データの価値を高めるだけではなく、新たな知見やデータの利用方法が提供者自身に還元されたり、研究コミュニティの拡大や新領域の開拓のきっかけになったりといったことも期待できる。無論、データを制限なく利用できるようにすることこそが正しいというわけではない。実際には、さまざまな段階の制限が設けられた提供方法（ライセンス）があり、試行錯誤が重ねられているところである。結局のところ、どのようにデータを公開・提供するのが最良であるかは、「何のためにデータやデータベースを公開

するのか」という、提供者自身の考えに帰着する問題である。その考えに基づきつつ、データの性質や利用者のニーズの変化、および、予想されるメリットやデメリットを見極めながら、随時、判断を重ねていく必要がある。

日文研が数多くのデータベースを公開していることは周知のとおりである。それぞれのデータベースで検索方法やデータの提示方法に工夫がなされており、多くの方にご利用いただいている。データそのものを利活用しようという新たな流れの中にあって、これらの有用な研究資源を今後どのように提供し、新たな可能性を模索してゆくのか、今まさに議論が始まろうとしているところである。

註

一　「トリプル・アイ・エフ」と呼称。

二　HuTime Project『HuTime Calendrical Period Resources』
http://datetime.hutime.org/

三　人文学オープンデータ共同利用センター『くずし字チャレンジ！』
http://codh.rois.ac.jp/kuzushiji-challenge/

（国際日本文化研究センター教授）